Diversity and Similarity Indices in Volunteered Geographic Information

Peter Mooney*[1]

[1]Department of Computer Science, Maynooth University, Co. Kildare. Ireland

*Email: peter.mooney@nuim.ie

Abstract

Volunteered Geographic Information (VGI) projects create various sizes of virtual and physical communities both online and offline. While there has been sustained interested in the quality of the data and information produced by these communities there are less results and analysis of the communities themselves. One of the key reasons for this is that there is often a lack of demographic information available for the participants in the communities. Therefore quantitative analysis is difficult. In this paper we introduce a novel and new approach to analysis of the communities in VGI using diversity, biotic and similarity indices from biological systems and aquatic ecosystems. These indices have been used for decades in these areas and our paper provides a preliminary exploration into how these indices might help us to better understand the structure and composition of VGI communities. We use OpenStreetMap (OSM) for the purposes of example.

Keywords: Volunteered Geographic Information, OpenStreetMap, Diveristy, Communities, Community Structure

1 Introduction and Motivation

OpenStreetMap (OSM) is probably the best known Volunteered Geographic Information (VGI) project on the Internet today. At the time of writing OSM is over 13 years in existence. However, despite extensive research interest in the OSM project over the past decade, little is still known about the characteristics and structures of the communities of volunteers who contribute to OSM on a daily basis. In April 2017 there are over 3.6 million registered contributors to OSM[1]. Several authors have quantitatively demonstrated that there is significant contribution inequality amongst the registered contributors to OSM. Work by Mooney and Corcoran (2012); Neis and Zipf (2012); Levin et al. (2017) indicate that a small number of contributors carry out and undertake the vast amount of work in the OSM mapping effort. Other authors such as Lin (2014) have considered the qualitative assessment of understanding the communities of VGI and OSM. The question addressed in this extended abstract paper is as follows: Is it possible to apply community structure metrics or indices from other scientific domains in order to provide quantitative understanding of VGI communities? The case-study is OSM. At the time of writing it is difficult to obtain quantitative demographic information about contributors to OSM. Subsequently different approaches to these

*

[1]http://osmstats.neis-one.org/

types of assessments must be considered (Barron et al., 2014). Fortunately the history of all edits to OSM is available for access, with contributor information on the edit actions performed.

The inspiration for this paper comes from Washington (1984), a seminal paper in the area of biological ecosystems, where the author outlines a large number of diversity, biotic and similarity indices for use on understanding the structure and composition of aquatic ecosystems. In order to try to understand the community structures of VGI data we have attempted to apply some of these indices on OSM for West Yorkshire. However, as we shall show we believe that this approach is applicable to other areas and regions in OSM and other VGI projects. This is particularly novel as these indices are normally only used by the biodiversity or aquatic ecosystems community. Washington (1984) motivates the use of these indices for understanding assemblages of species and their changes in abundance which has always been of interest to ecologists. Two approaches are used. Diversity indices attempt to combine data on abundance within species in a community into a single number. Similarity indices then take two samples and compare them using some measure of complement or dissimilarity. These indices are normally calculated for a given taxonomic grouping. In the case of OSM these groupings could be contributors, data objects, edit patterns, etc. Washington (1984) states that the indices should be applied to one geographical area.

2 Case-study: OpenStreetMap in West Yorkshire

For our experimental analysis we used the OSM history for West Yorkshire (OSM, 2017) which includes the city of Leeds, the location of Geocomputation 2017. The first recorded edit is from March 2006. There are $326,845$ distinct ways (polygons, polylines) in this dataset. There are $587,897$ way versions. There are $1,597$ distinct contributors to all of the ways in West Yorkshire from this time until March 25^{th} 2017. Of these contributors 193 of them are responsible for 95% of the mapping activity and editing corresponding to ways in the OSM database for West Yorkshire. The 10 most frequent editors are responsible for 55% of all mapping and editing efforts. Approximately 30% of contributors to the West Yorkshire OSM database have actually made 3 edits or less. There are many reasons for these inequalities. Some of the contributors may be actually very active in other areas of the United Kingdom or beyond. Some contributors may indeed be simply those to tried out OSM in the region but did not return. Authors such as Haklay (2013) have considered these mapping inequalities in greater detail.

There are just over $1,000$ contributors who were the inital creators of ways in West Yorkshire. The 10 most frequent editors are responsible for the creation of 75% of all ways in West Yorkshire. Overall these statistics give us a fuzzy picture of what is really happening in this OSM region. We can see that the region has a number of 'crazy mappers' or 'senior mappers' as Neis and Zipf (2012) labels them. However, these types of statistics do not allow us to understand how the 'crowd' of contributors to OSM are acting as a community.

2.1 Diversity, biotic and similarity indices

In this section we briefly outline the selection of indices from Washington (1984) which we apply to the OSM case-study data. Some notation is as follows. S is the total number of species. For our analysis we have to decide what a species actually is. N is the total number of individuals in the population or community. n is the number of individuals in a sample from the population or community. p_i is the fraction $\frac{n_i}{n}$ of a sample of individuals belonging to a species i

$$D = \sum_{i=1}^{S} \frac{n_i(n_i - 1)}{n(n-1)} \tag{1}$$

$$D = \frac{S-1}{\ln N} \tag{2}$$

$$H = -\sum_{i=1}^{S} \frac{n_i}{n} \ln \frac{ni}{n} \tag{3}$$

$$PIE = (\frac{N}{N-1})(1 - \sum_{i=1}^{S} p_i^2) \tag{4}$$

$$M = \frac{n - \sqrt{\sum_{i=1}^{S} n_i^2}}{n - \sqrt{n}} \tag{5}$$

In Equation 1 Simpson's D diversity index is given. This is the probability that two individuals chosen at random and independently from the population will be found to belong to the same group. This is one of the simplest approaches to diversity measurement. Rare species make little change to the community with more weighting given to abundant species. A value of 0 is infinitely diverse with 1 meaning no diversity amongst species.

In Equation 2 Margalef's D is based on the presumed linear relation between the number of species and the logarithm of the area of the number of individuals. This relates to the species richness.

Shannon's index is shown in Equation 3 and is one of the best known diversity indices. Washington (1984) calls it the 'magic bullet'. It takes its basis from Shannon's work in Information Theory. It has been used extensively throughout the biological world (Danilov and Ekelund, 1999). This measures the heterogeneity of the population by considering if two randomly sampled individuals are actually different species. When all species are equally common, all Shannon index has a maximum value $\ln(S)$.

In Equation 4 Hurlbert's PIE considers the concept of interspecies encounters and how each individual in a community can encounter or interact with every other individual in the community or population. Hurlbert's PIE ranges from 0 uneven to 1 even.

Equation 5 is known as McIntosh's "ecological distance" indicator. It refers to a measure of the ecological relationship suggested by the resemblance or similarity of two communities (Bandeira et al., 2013). This also ranges from 0 to 1 with the baseline 0 taken to establish the situtation where there are no individuals up to the point where every individual is a different a different species in the community.

2.2 Case-Study: Community Analysis in OSM

To apply the indices from Section 2.1 we need to automatically extract our synthetic communities containing distinct 'species' from the OSM data. This requires analysis of the OSM history data. We decided to consider three communities which we have developed ourselves from previous work in Mooney and Corcoran (2014):

1. **C1: Way Contribution Community**: We consider the grouping of contributors to OSM based on the total number of edits to ways in the OSM region. We created a community with 7 species based on the distribution of their overall contribution of edits and creation of ways. On species in this community is represented by contributors having contributed 10 or less edits to OSM in the chosen region. Another species in this community is represented by the very high frequency contributors who have contributed $5,000$ or more edits to OSM in the chosen region.
2. **C2: Way Creation Community**: We consider the grouping of contributors to OSM based on the total number of ways that they are responsible for creating in the OSM region. As previously we created a community with 7 species based on the distribution of their overall contribution of the creation of way objects in OSM for the chosen region.
3. **C3: Changeset Allocation Community**: We consider the grouping of contributors to OSM based on the number of distinct changesets that have in the OSM History. A changeset is a logical grouping of a group or collection of edits. Essentially they correspond to a collection of work carried out by a contributor. For this community, for example, we consider one species as those contirbutors with 5 or less changesets.

The number of species in each community can be easily changed and the indices in Section 2.1 can accomodate communities with different numbers of species.

Table 1: Results of application of the indices in Section 2.1 on three different communities

Index	**C1: Way Contributors**	**C2: Way Creators**	**C3: Changesets**
Simpsons D (Eqn: 1)	0.337	0.161	0.521
Margalef D (Eqn: 2)	0.949	0.949	0.949
Shannon (Eqn 3)	1.452	1.191	1.042
Shannon Max	1.945	1.945	1.945
Hurlberts PIE (Eqn 4)	0.662	0.838	0.478
McIntosh M (Eqn 5)	0.429	0.612	0.284

Table 1 indicates the results of the application of the indices in Section 2.1 to our three communities. There are a number of interesting observations which are obtained.

- Simpson's D (Eqn 1): This appears to indicate that there is more diversity amongst the contributors who actually created the initial representations of way objects in West Yorkshire than in the community of contributors which have edited and maintained the OSM database. Community C3 is less diverse given that changesets encapsulate groups of way edits and creations into one batch.
- Margalef D (Eqn: 2): This indicates that the number of species are the same in our three communities. The large value indicates that there is richness amongst the community.
- Shannon (Eqn 3): The maximum value of Shannon is supplied in the table. We see here that this index considers C1 to have more commonality amongst species than the other two communities. Again C3 has the lowest commonality.
- Hurlbert's PIE (Eqn 4): As before the changeset community C3 is the most unevenly structured community. However, in agreement with Simpson's D this index indicates that there is more eveness of distribution amongst those who are in community C2 than C1.

- McIntosh M (Eqn 5): Here we see this index indicating that there is more eveness amongst the members of C1 than C2.

3 Conclusions and Future Work

In this paper we considered the following research question: Is it possible to apply community structure metrics or indices from other scientific domains in order to provide quantitative understanding of VGI communities? We outlined some preliminary results in the analysis section in Section 2.2. We have shown that there are agreements and disagreements amongst this selection of indices. Community C3 appears to have a less evenly distributed spread of species. However C1 and C2 are more balanced. We believe that this is the basis for a more extensive analysis of community structure in VGI and OSM using this approach. This paper is the first known application of diversity, biotic and similarity indices from ecological aquatic ecosystems to VGI communities. As outlined in Washington (1984) many more indices exist. Bandeira et al. (2013) considers 7 indices and prove their mathematical convergence. Recent work by authors such as Chi et al. (2016); Arumugam et al. (2016) indicates the continued use of these approaches to community analysis in ecosystems.

One of the key challenges here is to create a suitable set of species, or contributor classes, from the OSM data. In this paper we considered contributors as being assigned to a particular class of species based on their editing activity in the OSM case-study region. We shall investigate more nuanced methods to develop classes of species in the data such as considering the types of edits made, types of syntatic and semantic contributions, etc. In the presentation at Geocomputation we shall provide an analysis of several different geographical regions in order to consider the similarity or otherwise of the distribution of species (contributors) in these areas.

4 Acknowledgements

The contributors to the OpenStreetMap project are acknowledged for their continued work on building one of the most extensive map databases in the world.

5 References

Arumugam, S., Sigamani, S., Samikannu, M. and Perumal, M. (2016), 'Assemblages of phytoplankton diversity in different zonation of Muthupet mangroves', *Regional Studies in Marine Science* **3**, 234–241.
URL: *http://www.sciencedirect.com/science/article/pii/S2352485515000729*

Bandeira, B., Jamet, J.-L., Jamet, D. and Ginoux, J.-M. (2013), 'Mathematical convergences of biodiversity indices', *Ecological Indicators* **29**, 522–528.
URL: *http://www.sciencedirect.com/science/article/pii/S1470160X13000605*

Barron, C., Neis, P. and Zipf, A. (2014), 'A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis', *Transactions in GIS* pp. 26–39.

Chi, S., Zheng, J., Zhao, X., Dong, F. and Hu, J. (2016), 'Macroinvertebrate communities and the relationships with biotic factors in river-connected lakes in the lower reaches of Yangtze River,

China', *Environmental Monitoring and Assessment* **188**(10), 577.
URL: *https://link.springer.com/article/10.1007/s10661-016-5602-y*

Danilov, R. and Ekelund, N. G. A. (1999), 'The efficiency of seven diversity and one similarity indices based on phytoplankton data for assessing the level of eutrophication in lakes in central Sweden', *Science of The Total Environment* **234**(13), 15–23.
URL: *http://www.sciencedirect.com/science/article/pii/S0048969799001631*

Haklay, M. (2013), 'Neogeography and the delusion of democratisation', *Environment and Planning A* **45**(1), 55–69.
URL: *http://www.envplan.com.jproxy.nuim.ie/abstract.cgi?id=a45184*

Levin, N., Lechner, A. M. and Brown, G. (2017), 'An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas', *Applied Geography* **79**, 115 – 126.
URL: *http://www.sciencedirect.com/science/article/pii/S0143622816308268*

Lin, W. (2014), 'Revealing the making of OpenStreetMap: A limited account', *The Canadian Geographer / Le Gographe canadien* pp. n/a–n/a.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/cag.12137/abstract*

Mooney, P. and Corcoran, P. (2012), 'The Annotation Process in OpenStreetMap', *Transactions in GIS* **16**(4), 561–579.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2012.01306.x/abstract*

Mooney, P. and Corcoran, P. (2014), 'Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors', *Transactions in GIS* **18**(5), 633–659.
URL: *http://onlinelibrary.wiley.com/doi/10.1111/tgis.12051/abstract*

Neis, P. and Zipf, A. (2012), 'Analyzing the Contributor Activity of a Volunteered Geographic Information Project The Case of OpenStreetMap', *ISPRS International Journal of Geo-Information* **1**(2), 146–165.
URL: *http://www.mdpi.com/2220-9964/1/2/146*

OSM, W. Y. (2017), 'The OSM relation for west yorkshire, uk'.
URL: *http://www.openstreetmap.org/relation/88079*

Washington, H. (1984), 'Diversity, biotic and similarity indices', *Water Research* **18**(6), 653 – 694.
URL: *http://www.sciencedirect.com/science/article/pii/0043135484901647*